PDSF
NERSC's Production Linux Cluster

Craig E. Tull

HCG/NERSC/LBNL

MRC Workshop

LBNL *- March 26, 2002*

Outline

- **People: Shane Canon (Lead), Cary Whitney, Tom Langley, Iwona Sakredja (Support)**
  (Tom Davis, Tina Declerk, John Milford, others)

- **Present**
  - —**What is PDSF?**
  - —**Scale, HW & SW Architecture, Business & Service Models, Science Projects**

- **Past**
  - —**Where did PDSF come from?**
  - —**Origins, Design, Funding Agreement**

- **Future**
  - —**How does PDSF relate to the MRC initiative?**

PDSF - Production Cluster

- **PDSF - Parallel Distributed Systems Facility**
  - **HENP community**
    - **Specialized needs/Specialized requirements**
- **Our mission is to provide the most effective distributed computer cluster possible that is suitable for experimental HENP applications.**
- **Architecture tuned for "embarrassingly parallel" applications**
- **AFS access, and access to HPSS for mass storage**
- **High speed (Gigabit Ethernet) access to HPSS system and to Internet Gateway**
- **http://pdsf.nersc.gov/**

PDSF Photo

PDSF Overview

PDSF- Jan 2002

NERSC & Internet

Extreme 7i

Summary

207 Compute Nodes
390 Processors
 35 TB Total Disk
360 GB Total Memory
 75 Gigabit Nodes

Cisco3548

Sun E450
1.5 TB Disk
4 GB RAM

Near EOL hardware
21 Computenodes
7 Diskvaults

Administrative
  Servers
Developement Nodes
Console Servers

Small Diskvaults
3 Nodes
Dual 677 Mhz PIII
300 GB RAID0

Extreme 7i

Extreme 7i

Extreme 7i

Extreme 7i

Cisco3548

High Bandwidth Nodes
42 Nodes
Dual 1GHz PIII
Gigabit
2 GB RAM
300 GB Disk

Disk Vaults
26 Nodes
Dual 1GHz PIII
Gigabit
2 GB RAM
0.5 TB RAID5

Raidzones
6 Nodes
Dual 677 MHz PIII
Gigabit
1 GB RAM
1 TB RAID5

Compute Nodes
4 Racks
1 Switch/Rack
32 Nodes/Rack
88 Dual 677 MHz PIII
34 Dual 1GHz PIII
100baseT
2 GB RAM
>40GB Disk

LAWRENCE BERKELEY NATIONAL LABORATORY

Component HW Architecture

- **By minimizing diversity within the system, we maximize uptime and minimize sys-admin overhead.**
- **Buying Computers by the Slice (Plant & <u>Prune</u>)**
  - **Buy large, homogeneous batches of HW.**
  - **Each large purchase of HW can be managed as a "single unit" composed of interchangable parts.**
  - **Each slice has limited lifespan - NO MAINT. $**
- **Critical vs. non-Critical Resources**
  - **Non-critical compute nodes can fail without stopping analysis.**
  - **Critical data vaults can be trivially interchanged (with transfer of disks) with compute nodes.**
- **Uniform environment means that software & security problems can be solved by "Reformat & Reload".**

- **Interactive**
  - —**8 Intel Linux (RedHat 6.2)**
  - —**More memory, Fast, interactive logon, serve batch jobs when idle**
- **Batch (Normal & High Bandwidth)**
  - —**~400 Intel Linux CPUs (RedHat 6.2)**
  - —**LSF: Short, Medium, Long, & Custom Queues**
- **Data Vault**
  - —**Large Shared (NFS) Disk Arrays - 25 TB**
  - —**High Network (GigE) Connectivity to compute nodes & HPSS, Data-Transfer Jobs Only**
- **Administrative**
  - —**Home diskspace server, AFS servers, License servers, Database servers, time servers, etc.**

PDSF Global SW

- **All SW necessary to HENP data analysis & simulation is available and maintained at current revs**

- **Solaris Software**
  - **AFS, CERN libs, CVS, Modules, Objectivity, Omnibroker, Orbix, HPSS pFTP, ssh, LSF, PVM, Framereader, Sun Workshop Suite, Sun's Java Dev. Toolkit, Veritas, Python, etc.**

- **Linux Software**
  - **AFS, CERN libs, CVS, Modules, Omnibroker, ssh, egcs, KAI C++, Portland Group F77/F90/HPF/C/C++, LSF, HPSS pFTP, Objectivity, ROOT, Python, etc.**

- **Specialized Software (Experiment Maintained)**
 - **ATLAS, CDF, D0, DPSS, E895, STAR, etc.**

Experiment/Project SW

- **Principal: Allow diverse groups/projects to operate on all nodes without interfering with others.**
- **Modules:**
  - **Allows individuals/groups to chose appropriate SW versions at login (Version migration dictated by experiment, not system.).**
- **Site independence:**
  - **PDSF personnel have been very active in helping "portify" code (STAR, ATLAS, CDF, ALICE).**
    - **Direct benefit to project Regional Centers & institutions.**
- **Specific kernel/libc dependency of project SW is only case where interference is an issue (None now.).**
- **LSF extensible to allow incompatible differences.**

Batch Queuing System
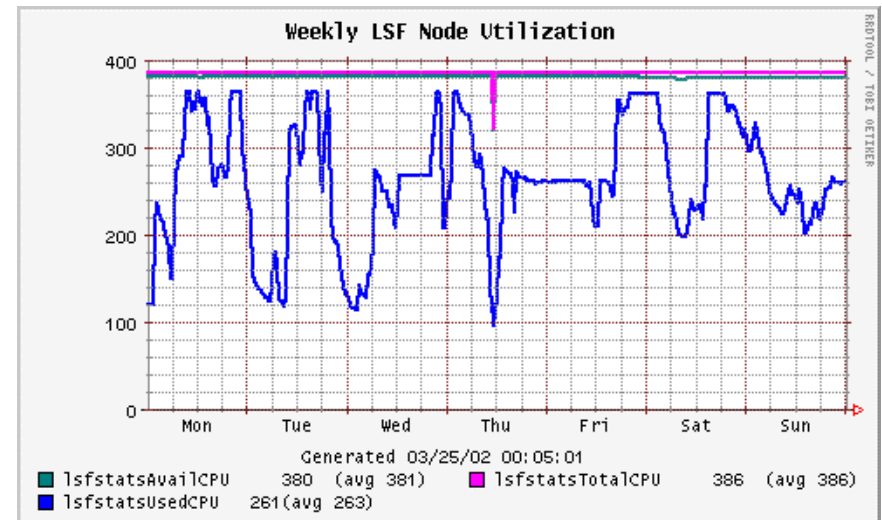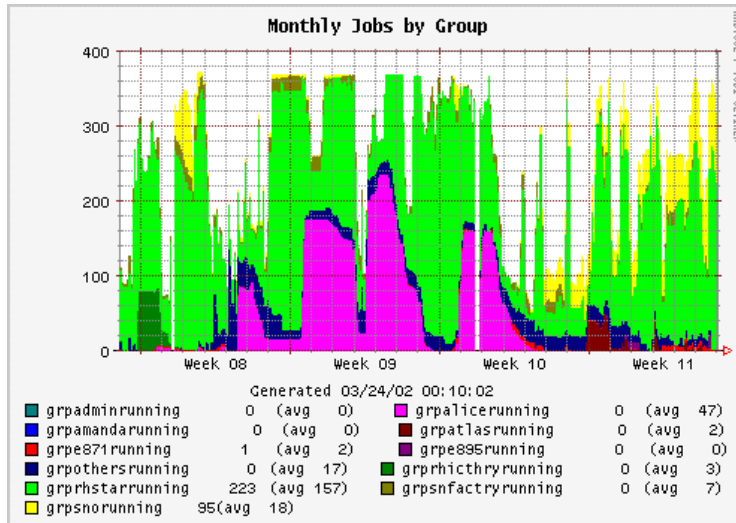
- **LSF 4.x - Load Sharing Facility**
  - **Solaris & Linux**
  - **Tremendous leverage from NERSC (158 "free", aggressive license negotiations Þ price savings)**
  - **Very good user & admin experiences**
  - **Fair share policy in use**
  - **Can easily sustain >95% load.**

Administrative SW/Tools

- **Monitoring tools - Batch usage, node & disk health, etc.**
  - **developed at PDSF to insure smooth operation & assure contributing clients**
- **HW tracking & location (mySQL + ZOPE)**
  - **800 drives, ~300 boxes, HW failures/repairs, etc.**
  - **developed at PDSF out of absolute necessity**
- **System & Package consistency**
  - **developed at PDSF**
- **System installation**
  - **kickstart - ~3 min.s/node**
- **System Security - No Known Security Breaches**
  - **TCP Wrappers & ipchains, NERSC Monitoring, no clear-text passwords, security patches, crack, etc.**

PDSF Business Model

- **Users contribute directly to cluster through hardware purchases. The size of the contribution determines the fraction of resources that are guaranteed.**

- **NERSC provides facilities and administrative resources (up to pre-agreed limit).**

- **User share of resource guaranteed at 100% for 2 years (warranty), then depreciates 25% per year (hardware lifespan).**

- **PDSF scale has reached the point where some FTE resources must be funded.**
  - **#Admins Ü Size of System (eg. box count)**
  - **#Support Ü Size of User Community (eg. #groups & #users & diversity)**

PDSF Service Model

- **Less than 24/7, but more than Best Effort.**
- **Support (1 FTE):**
  - —**USG supported web-based trouble tickets.**
  - —**Response during business hours.**
  - —**Performance matrix.**
  - —**Huge load right now Ü Active, Large community**
- Admin (3 FTE):
  - —**NERSC operations monitoring (24/7)**
  - —**Critical vs. non-critical resources**
  - —**Non-critical: (eg. batch nodes) Best effort**
  - —**Critical: (eg. servers, DVs) Fastest response**

History of PDSF Hardware

- **Arrived from SSC (RIP), May 1997**
- **October '97 - "Free" HW from SSC**
  - **32 SUN Sparc 10, 32 HP 735**
  - **2 SGI data vaults**
- **1998 - NERSC seeding & initial NSD updates**
  - **added 12 Intel (E895), SUN E450, 8 dual-cpu Intel (NSD/STAR), 16 Intel (NERSC), 500 GB network disk (NERSC)**
  - **subtracted SUN, HP, 160 GB SGI data vaults**
- **1999 - Present - Full Plant & Prune**
  - **HENP Contributions: STAR, E871, SNO, ATLAS, CDF, E891, others**
- **March 2002**
  - **240 Intel Compute Nodes (390 CPUs)**
  - **8 Intel Interactive Nodes (dual 996MHz PIII, 2GB RAM, 55GB scratch)**
  - **49 Data Vaults: 25 TB of shared disk**
  - **Totals: 570K MIPS, 35 TB disk**

PDSF Growth

- **Future**
  - **Continued support of STAR & HENP**
  - **Continue to seek out new groups**
  - **Look for interest outside of HENP community**
    - **Primarily Serial workloads**
    - **Stress benefits of using a shared resource**
    - **Let scientist focus on science and system administrator run computer systems**
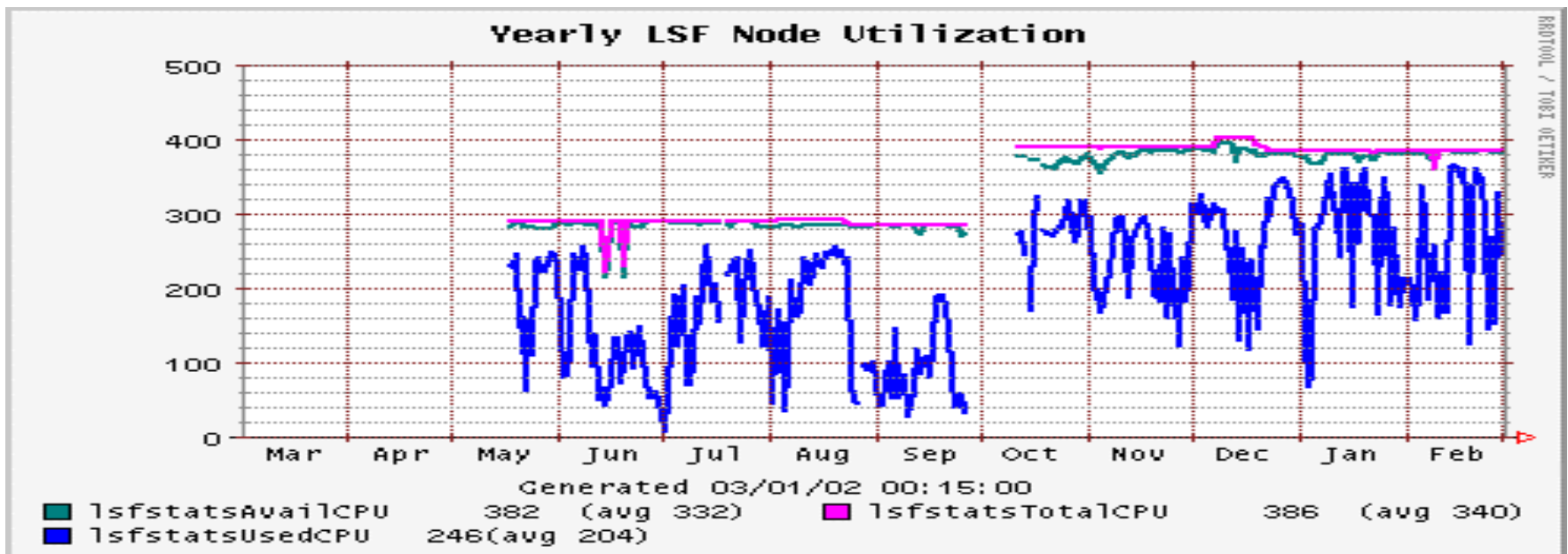
PDSF Major/Active Users

- **Collider Facilities**
  - **STAR/RHIC - Largest user**
  - **CDF/FNAL**
  - **ATLAS/CERN**
  - **E871/FNAL**
- **Neutrino Experiments**
  - **Sudbury Neutrino Observatory (SNO)**
  - **KamLAND**
- **Astrophysics**
  - **Deep Search**
  - **Super Nova Factory**
  - **Large Scale Structure**
  - **AMANDA**

Solenoidal Tracking At RHIC (STAR)

- **Experiment at the RHIC accelerator in BNL**
- **Over 400 scientists and engineers from 33 institutions in 7 countries**
- **PDSF primarily intended to handle simulation workload**
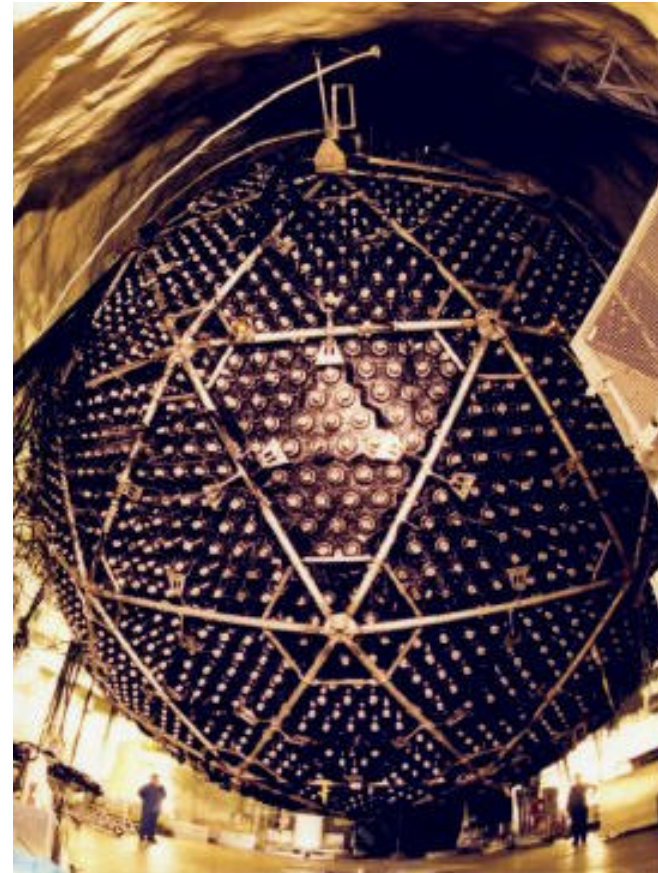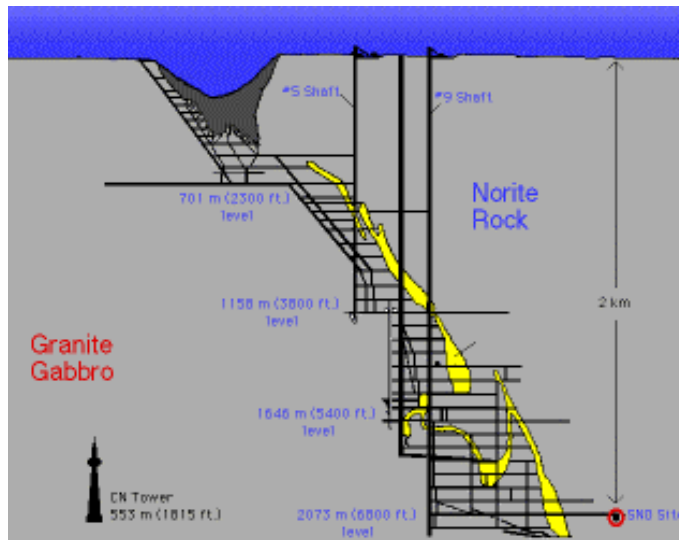- **PDSF has increasingly been used for general analysis**
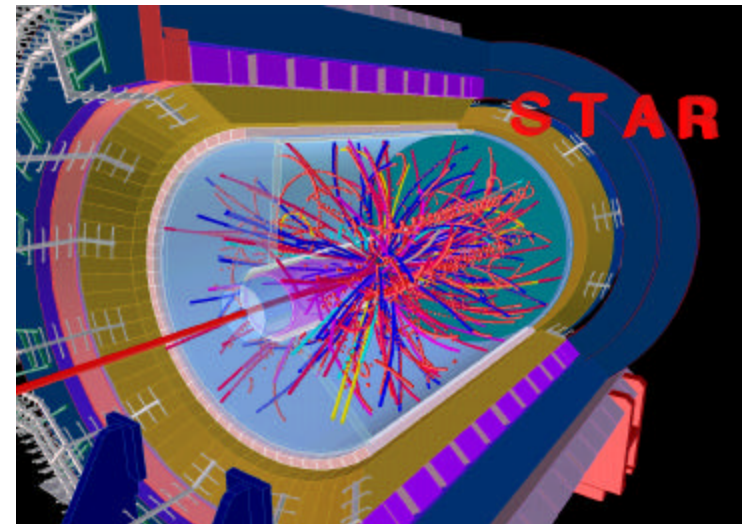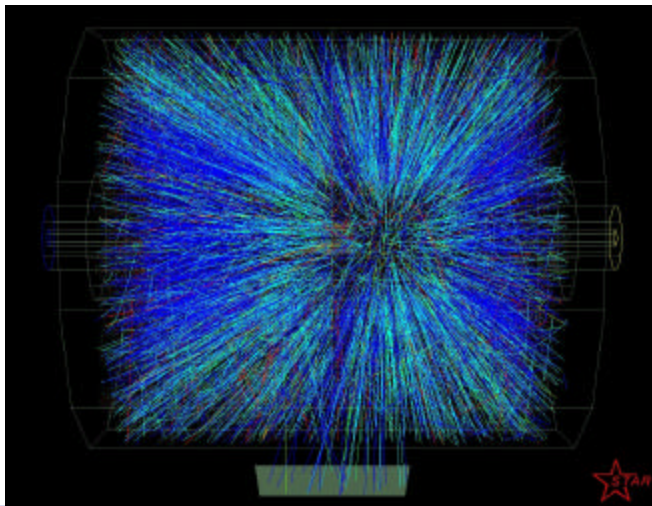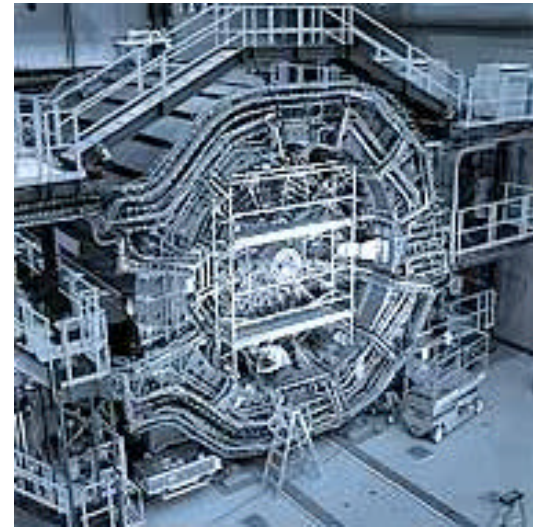
Sudbury Neutrino Observatory (SNO)

- **Located in a mine in Ontario Canada**
- **Heavy water neutrino detector**
- **SNO has over 100 collaborators at 11 institutions**

Results from SNO

- **Recently confirmed results from Super-K that neutrinos have mass.**
- **PDSF specifically mentioned in results**

Measurement of the rate of $\nu_e + d \to p + p + e^-$ interactions produced by ^{8}B solar neutrinos at the Sudbury Neutrino Observatory

Q.R. Ahmad[15], R.C. Allen[11], T.C. Andersen[12], J.D. Anglin[7], G. Bühler[11], J.C. Barton[13‡], E.W. Beier[14], M. Bercovitch[7], J. Bigu[4], S. Biller[13], R.A. Black[13], I. Blevis[3], R.J. Boardman[13], J. Boger[2], E. Bonvin[9], M.G. Boulay[9], M.G. Bowler[13], T.J. Bowles[6], S.J. Brice[6,13], M.C. Browne[15], T.V. Bullard[15], T.H. Burritt[15,6], K. Cameron[12], J. Cameron[13], Y.D. Chan[5], M. Chen[9], H.H. Chen[11*], X. Chen[5,13], M.C. Chon[12], B.T. Cleveland[13], E.T.H. Clifford[9,1], J.H.M. Cowan[4], D.F. Cowen[14], G.A. Cox[15], Y. Dai[9], X. Dai[13], F. Dalnoki-Veress[3], W.F. Davidson[7], P.J. Doe[15,11,6], G. Doucas[13], M.R. Dragowsky[6,5], C.A. Duba[15], F.A. Duncan[9], J. Dunmore[13], E.D. Earle[9,1], S.R. Elliott[15,6], H.C. Evans[9], G.T. Ewan[9], J. Farine[3], H. Fergani[13], A.P. Ferraris[13], R.J. Ford[9], M.M. Fowler[6], K. Frame[13], E.D. Frank[14], W. Frati[14], J.V. Germani[15,6], S. Gil[10], A. Goldschmidt[6], D.R. Grant[3], R.L. Hahn[2], A.L. Hallin[9], E.D. Hallman[4], A. Hamer[6,9], A.A. Hamian[15], R.U. Haq[4], C.K. Hargrove[3], P.J. Harvey[9], R. Hazama[15], R. Heaton[9], K.M. Heeger[15], W.J. Heintzelman[14], J. Heise[10], R.L. Helmer[10‡], J.D. Hepburn[9,1], H. Heron[13], J. Hewett[4], A. Hime[6], M. Howe[15], J.G. Hykawy[4], M.C.P. Isaac[5], P. Jagam[12], N.A. Jelley[13], C. Jillings[9], G. Jonkmans[4,1], J. Karn[12], P.T. Keener[14], K. Kirch[6], J.R. Klein[14], A.B. Knox[13], R.J. Komar[10,9], R. Kouzes[8], T. Kutter[10], C.C.M. Kyba[14], J. Law[12], I.T. Lawson[12], M. Lay[13], H.W. Lee[9], K.T. Lesko[5], J.R. Leslie[9], I. Levine[3], W. Locke[13], M.M. Lowry[8], S. Luoma[4], J. Lyon[13], S. Majerus[13], H.B. Mak[9], A.D. Marino[5], N. McCauley[13], A.B. McDonald[9,8], D.S. McDonald[14], K. McFarlane[3], G. McGregor[13], W. McLatchie[9], R. Meijer Drees[15], H. Mes[3], C. Mifflin[3], G.G. Miller[6], G. Milton[1], B.A. Moffat[9], M. Moorhead[13,5], C.W. Nally[10], M.S. Neubauer[14],

PDSF Job Matching

- **Linux clusters are already playing a large role in HENP and Physics simulations and analysis.**
- **Beowulf systems may be in-expensive, but can require lots of time to administer**
- **Serial Jobs:**
  - **—Perfect match (up to ~2GB RAM+2GB SWAP)**
- **MPI Jobs:**
  - **—Some projects using MPI (Large Scale Structures, Deep Search).**
  - **—Not low latency, small messages**
- **"Real" MPP (eg. Myricom)**
  - **—Not currently possible. Could be done, but entails significant investment of time & money.**
  - **—LSF handles MPP jobs (already configured).**

How does PDSF relate to MRC?

- **Emulation or Expansion ("Model vs. Real Machine")**
  - —**If job characteristics & resources match.**
- **Emulation:**
  - —**Adopt appropriate elements of HW & SW Arch.s, Business & Service Models.**
  - —**Steal appropriate tools.**
  - —**Customize to eg. non-PDSF like job load.**
- **Expansion:**
  - —**Directly contribute to PDSF resource.**
  - —**Pros:**
    - **Faster ramp-up, stable environment, try before bye, larger pool of resources (better utilization), leverage existing expertise/infrastructure**
  - —**Cons:**
    - **Not currently MPP tuned, ownership issues.**